03-31-00

A

FISH & RICHARDSON P.C.

225 Franklin Street
Boston, Massachusetts
02110-2804

Telephone
617 542-5070

Facsimile
617 542-8906

Web Site
www.fr.com

March 30, 2000

Attorney Docket No.: 07064-010001

Box Patent Application
Assistant Commissioner for Patents
Washington, DC 20231

Presented for filing is a new original patent application of:

Applicant: SAMIR KUMAR BRAHMACHARI AND DEBASIS DASH

Title: A COMPUTER BASED METHOD FOR IDENTIFYING PEPTIDES USEFUL AS DRUG TARGETS

Enclosed are the following papers, including those required to receive a filing date under 37 CFR 1.53(b):

	Pages
Specification	17
Claims	4
Declaration	[To be Filed at a Later Date]
Abstract	1
Drawing(s)	4

Enclosures:
— Postcard.

Basic filing fee	$690
Total claims in excess of 20 times $18	$0
Independent claims in excess of 3 times $78	$78
Fee for multiple dependent claims	$260
Total filing fee:	$1028

CERTIFICATE OF MAILING BY EXPRESS MAIL

Express Mail Label No. EL224670093US

I hereby certify under 37 CFR §1.10 that this correspondence is being deposited with the United States Postal Service as Express Mail Post Office to Addressee with sufficient postage on the date indicated below and is addressed to the Assistant Commissioner for Patents, Washington, D C 20231.

March 30, 2000
Date of Deposit

Samantha Bell
Signature

Samantha Bell
Typed or Printed Name of Person Signing Certificate

A check for the filing fee is enclosed. Please apply any other required fees or any credits to deposit account 06-1050, referencing the attorney docket number shown above.

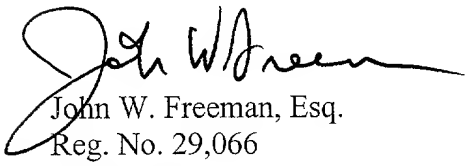
If this application is found to be incomplete, or if a telephone conference would otherwise be helpful, please call the undersigned at (617) 542-5070.

Kindly acknowledge receipt of this application by returning the enclosed postcard.

Please send all correspondence to:

JOHN W. FREEMAN, ESQ.
Fish & Richardson P.C.
225 Franklin Street
Boston, MA 02110-2804

Respectfully submitted,



John W. Freeman, Esq.
Reg. No. 29,066
Enclosures
JWF/lxv
20044880.doc

APPLICATION

FOR

UNITED STATES LETTERS PATENT

TITLE: A COMPUTER BASED METHOD FOR IDENTIFYING PEPTIDES USEFUL AS DRUG TARGETS

APPLICANT: SAMIR KUMAR BRAHMACHARI AND DEBASIS DASH

CERTIFICATE OF MAILING BY EXPRESS MAIL

Express Mail Label No EL224670093US

I hereby certify under 37 CFR §1.10 that this correspondence is being deposited with the United States Postal Service as Express Mail Post Office to Addressee with sufficient postage on the date indicated below and is addressed to the Assistant Commissioner for Patents, Washington, D.C. 20231.

March 30, 2000
Date of Deposit

Samantha Bell
Signature

Samantha Bell
Typed or Printed Name of Person Signing Certificate

Field of the invention:

This invention relates to a computer-based method for identifying peptides useful as drug targets. More particularly this invention relates to a method for identification of invariant peptide motifs in protein sequence data of various organisms useful as potential drug targets. This invention further provides a method for assignment of function to hypothetical Open Reading Frames (proteins) of unknown function through exact amino acid sequence identity signature.

This invention provides a novel approach for identifying structural and functional signatures of conserved invariant amino acid sequences of proteins that can serve as potential candidates for drug targets. Emergence of drug resistant strains has necessitated identification of new drugs and drug targets. Unique invariant peptide motifs present in the proteins of pathogen but absent in the proteins of host indicate potential drug targets. The invention also provides a method for genome wise comparison of large number of protein sequences simultaneously. Yet another utility is for identifying peptide sequences useful for specific diagnosis of infections.

Background of the invention:

It is known that most of the drugs that are available today to cure infections bind to specific protein target molecules in the cell of the causative organism e.g., several antibiotics are known to disrupt the function of ribosomes so that the protein translation is affected. In these cases it has been found that the drugs either bind to the ribosomal RNA directly or RNA protein complexes (Wimberly et al, 1999). Chemical probing experiments have revealed that the drug binds to certain nucleotide sequences of ribosomal RNA that are 'invariant' in structurally analogous regions in different organisms (Porse and Garrett, 1999). The other class of drugs serves to block other functions such as transcription (Cutler et al, 1999) or fatty acid synthesis in the bacterial cell (McCafferty et al., 1999).

Recently, several drug resistant strains (Ghannoum & Rice, 1999) of pathogenic bacteria have emerged that renders the current treatment procedures ineffective in curing infections due to bacterial pathogens. This necessitates the identification of new drug targets and the corresponding drugs. For this purpose, the availability of complete genome sequences from various microbes offers us an opportunity to analyze all the proteins encoded in a given genome. Since most drugs

known today target proteins, it is likely that analyzing all the proteins in a given bacterium may provide new valid drug targets.

The knowledge of conserved invariant sequences in a protein can be useful in understanding certain features of a protein's architecture, such as buried versus exposed location of a segment or the presence of specific secondary structural elements (Rooman and Wodak, 1988, Presnell et al., 1992). The protein's functional role is the most important aspect of conserved invariant sequences. Methods of usual sequence analysis include BLAST (Altschul et al., 1990), and FASTA (Wilbur and Lipman, 1983). These methods carryout sequence alignments whose quality is evaluated using an amino acid substitution matrix. Statistical calculations are performed and the results are output in a ranked manner, with the best similar sequence ranking first. However, these methods are not designed to do a genome-wise comparison simultaneously to identify invariant sequence motifs that are of particular importance in this work.

In order to compare each protein of one organism with all other proteins of several other organisms, either one has to use BLAST one by one or a batch BLAST has to be used which is highly time consuming and therefore not practicable. Even if this were done, at the end of the exercise, one would obtain the overall similarity of a set of homologous proteins and alignments.

The problem with multiple sequence alignment is that it is biased to the selection of proteins. Only proteins that are functionally related will give a clear picture of any relationship between the selected proteins. Such procedures are labor intensive and time consuming and leads to results that need further processing and filtering. However, by these methods it is not possible to compare all proteins of several organisms and retrieve conserved invariant peptides.

The present invention provides a novel computer based method to look for invariant sequence motif that will lead to manifold usage as mentioned above and obviates the drawbacks listed above.

The applicants' approach is based on the paradigm that the invariant sequence motifs between the different bacterial proteins must be responsible for an important role for the structure and the function of the protein. Of the numerous ways by which drug targets can be identified, we have taken an approach based on comparative & structural genomics. In this case, the invariant sequence motifs may be either directly or indirectly involved in the function of the subject protein molecule. This approach is derived from the concept that invariant sequence motifs that have remained unchanged across bacteria that are related either distantly or closely should have evolved

a unique structural feature that can not be compromised. Indeed, it is even possible that the so-called conservative substitutions are also not tolerated in these invariant sequence motifs. To this end, we have identified several invariant peptide motifs by direct sequence comparison between various bacterial genomes without any *a priori* assumptions. This purely unbiased and unassumed way of studying the sequences has the benefit of revealing unidentified sequence properties in the various genomes.

Since the invariant sequence motifs may be important for the function of the subject protein molecule, we aim to develop these peptide motifs as potential broad-spectrum antibacterial drug targets. It is probable that a small molecule that can bind specifically to these invariant sequences may cause disruption of function of the subject protein molecule. It is envisaged that this *in silico* approach will provide new leads for experimental validation to derive functions from protein sequences existing in the available databases.

Objects of the invention:

The main object of the present invention is to provide a method for genome-wise protein sequence comparison of several organisms and identification of invariant conserved peptides.

Another object of the present invention relates to a novel computer based method for performing genome-wise comparison of several organisms, wherein the said computational method involves creation of peptide libraries from protein sequences of several organisms and subsequent comparison leading to identification of conserved invariant peptide motifs.

Yet another object of the present invention relates to providing a method useful for identification of potential drug targets and can serve as drug screen for broad spectrum antibacterials as well as for specific diagnosis of infection.

Another object of the present invention is to assign suitable function to proteins of yet unknown functions.

Yet another object is to provide a computational method incorporating the invariant peptides or their analogs for identifying potential drug targets.

Summary of the Invention:

The applicants have invented a method to identify invariant peptide motifs, obtained from millions of peptides present in protein sequences of many organisms that has withstood natural selection. These sequences are thus structural determinants of proteins, which could be targeted or can be used as screen as target for drug discovery. These special invariant peptide signatures are also fund to be associated with special functional class of proteins.

The present method will also allow predicting toxicity, alternate target in host cell for drug targeted against a specific peptide motif of a pathogenic organism or any host protein target responsible for a disease process. The method could be extended with lower stringencies to larger number of proteins and also for eukaryotes and multicellular organisms.

Other and further aspects, features and advantages of the present invention will be apparent from the following description of the presently preferred embodiments of the invention given for the purpose of disclosures.

Brief description of the computer programs:

1. PEPLIB

Objective: To create peptide libraries of organisms from their FASTA format protein files. Thus overlapping peptides of user defined length are generated and then only non-redundant peptides are arranged alphabetically in the output file.

Programming language: PERL on IRIX platform.

2. PEPLIMP

Objective: This program compares the peptide libraries of organisms selected by the user and returns the peptides sequences that are common across the genomes.

Programming language: PERL on IRIX platform.

3. **PEPXTRACT**

Objective: This program takes peptide file as input, searches in the FASTA format protein files (pep files) and returns the details about the peptides. The details include the PID, location of the peptide in the protein, Organism name etc.

Programming language: PERL on IRIX platform.

4. **PEPSTITCH**

Objective: This program joins the peptides depending on certain fixed criteria (the two peptides should have the same PID and their locations should be adjacent) and removes overlappings and reports all the conserved invariant peptides.

Programming language: PERL on IRIX platform.

Details of the invention:

Theoretically speaking, though, a huge number of combinations are possible at amino acid level to form a peptide of a given length only a limited fraction has been observed in biological systems. Out of this limited fraction, only a few peptides remained invariant across the genomes of different organisms. In this work, we sought to answer the question pertaining to the nature of peptides that are invariant across all the pathogenic and nonpathogenic bacterial genome.

In the present invention it has been shown that a stretch of amino acid conservation in proteins of various organisms can provide accurate distinction between different classes of proteins. Generally, these proteins are identified as proteins having very basic function in the survival of the organism.

The protein sequences of several organisms were obtained computationally from the existing databases (NCBI, genbank/genomes/bacteria). These were then chopped computationally into peptide fragments of 'N' amino acid residues by a specially developed computer program PEPLIB. A library of peptides of length 'N' was created for all the proteins of each organism by sliding the window of length 'N' along the sequence by one residue at a time. The peptides thus obtained were computationally sorted in an alphabetical order according to single letter amino acid

code, and the redundancy was removed by deleting duplicated peptides. The peptide libraries of various organisms were then compared computationally to find out common peptides. The comparison was done using a specially developed computer program labeled PEPLIMP. The common peptides were located computationally in the original proteins using PEPXTRACT program and were subsequently labeled with their proteins of origin and location. These common peptides were backstitched computationally to form a long chain of common peptides. This was done using PEPSTICH program.

These fragments of common peptides thus obtained were termed as invariant peptides as they originated from functionally conserved proteins. All the conserved invariant peptides obtained from the same protein were then clustered into one group. The secondary structure of these peptides was validated from the protein crystal structure database namely Protein Data Bank (PDB).

Accordingly the invention provides a computer-based method for identifying invariant peptide motifs useful as drug targets wherein the said method comprises the steps of:

i) generating computationally overlapping peptide libraries from all the protein sequences of the selected organisms available at http://www.ncbi.nlm.nih.gov,

ii) sorting computationally the peptides of length 'N' obtained as above, alphabetically, according to single letter amino acid code,

iii) matching computationally common peptide sequences of the selected bacteria,

iv) locating computationally these common peptides in the original proteins and subsequently labeling them with their origin and location,

v) joining computationally the overlapping common peptides to obtain a long chain of invariant peptide sequences,

vi) annotating secondary structure of these conserved peptides from the crystal structure database,

vii) comparing pathogenic strain genomes against genomes of non-pathogenic strains and selecting the sequences not commonly conserved in these two groups,

viii) validating computationally the invariant sequence motifs as potential drug target sequence by searching for the given conserved sequences in the host genome and rejecting the ones present in the host genome.

In an embodiment to the present invention the length of the sliding window of length 'N' may range from 4 to any length of amino acid residues.

In another embodiment to the present invention the protein sequence data may be taken from any organism but not specifically limited to microbes such as *Mycoplasma pneumoniae*, *Helicobacter pylori*, *Hemophillus influenzae*, *Mycobacterium tuberculosis*, *Mycoplasma genitalium*, *Bacillus subtillis*, *Escherichia coli*.

In further embodiment the conserved peptide motifs as identified comprise:

1. AAQSIGEPGTQLT
2. AGDGTTTAT
3. AGRHGNKG
4. AHIDAGKTTT
5. CPIETPEG
6. DEPSIGLH
7. DEPTSALD
8. DEPTTALDVT
9. DHAGIATQ
10. DHPHGGGEG
11. DLGGGTFD
12. DVLDTWFSS
13. ERERGITI
14. ERGITITSAAT
15. ESRRIDNQLRGR
16. FSGGQRQR
17. GEPGVGKTA
18. GFDYLRDN
19. GHNLQEHS
20. GIDLGTTNS
21. GINLLREGLD
22. GIVGLPNVGKS
23. GKSSLLNA
24. GLTGRKIIVDTYG
25. GPPGTGKTLLA
26. GPPGVGKT
27. GSGKTTLL
28. GTRIFGPV
29. IDTPGHVDFT
30. IIAHIDHGKSTL
31. INGFGRIGR
32. IREGGRTVG
33. IVGESGSGKS
34. KFSTYATWWI
35. KMSKSKGN
36. KMSKSLGN
37. KNMITGAAQMDGAILVV
38. KPNSALRK
39. LFGGAGVGKTV
40. LGPSGCGK
41. LHAGGKFD
42. LIDEARTPLIISG
43. LLNRAPTLH
44. LPDKAIDLIDE
45. LPGKLADC
46. LSGGQQQR
47. MGHVDHGKT
48. NADFDGDQMAVH
49. NGAGKSTL
50. NLLGKRVD
51. NTDAEGRL
52. PSAVGYQPTLA
53. QRVAIARA
54. QRYKGLGEM
55. RDGLKPVHRR
56. SALDVSIQA
57. SGGLHGVG
58. SGSGKSSL
59. SGSGKSTL
60. SVFAGVGERTREGND
61. TGRTHQIRVH
62. TGVSGSGKS
63. TLSGGEAQRI
64. TNKYAEGYP
65. TPRSNPATY
66. VEGDSAGG
67. VRKRPGMYIG

In yet another embodiment to the present invention, the number of invariant peptides may vary according to the relatedness among the organisms and the number of organisms being compared.

In still another embodiment, the invariant sequences may belong to following proteins as available in the database http://www.ncbi.nlm.nih.gov wherein the said list of proteins comprise:

I DNA DIRECTED RNA POLYMERASE BETA CHAIN

II EXCINUCLEASE ABC SUBUNIT A

III EXCINUCLEASE ABC SUBUNIT B

IV DNA GYRASE SUBUNIT B

V ATP SYNTHASE BETA CHAIN

VI S-ADENOSYLMETHIONINE SYNTHETASE

VII GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE

VIII ELONGATION FACTOR G (EF-G)

IX ELONGATION FACTOR TU (EF-TU)

X 30S RIBOSOMAL PROTEIN S12

XI 50S RIBOSOMAL PROTEIN L12

XII 50S RIBOSOMAL PROTEIN L14

XIII VALYL tRNA SYNTHETASE (VALRS)

XIV CELL DIVISON PROTEIN FtSH HOMOLOG

XV DnaK PROTEIN (HSP70)

XVI GTP BINDING PROTEIN LepA

XVII TRANSPORTER

XVIII OLIGOPEPTIDE TRANSPORT ATP BINDING PROTEIN OPPF

In still another embodiment to the present invention, the said method of comparing the peptide libraries as given in step (iii) of method explained above is carried out by following the steps given in figure 1.

In yet another embodiment to the present invention, the said method of locating the common peptides in the original protein sequences as given in step (iv) method explained above is carried out by following the steps given in figure 2.

In another embodiment, the method of creating a common peptide of variable length after removing the overlappings as given in step (v) of method explained above is carried out by following the steps given in figure 3.

In another embodiment to the present invention, the microprocessor based system for performing the methods of the invention comprises:

i) means of determining the amino acid sequence window for creation of peptide library and subsequent sorting,

ii) means of comparing the peptide library,

iii) locating computationally these common peptides in the original proteins and subsequently labeling them with their origin and location,

iv) joining computationally the overlapping common peptides to obtain a long chain of invariant peptide sequences,

In another embodiment of the invention, the computer system for performing the methods of the invention comprises, a central processing unit, executing peptide library creating program (PEPLIB), peptide library matching program (PEPLIMP), peptide stitching program (PEPSTITCH), peptide extraction program (PEPXTRACT) wherein the said programs are all stored in a memory device accessed by the central processing unit connected to a display on which the central processing unit displays the screens of the above mentioned programs in response to user inputs with a user interface device.

In yet another embodiment to the present invention, the method for assigning function to a protein of unknown function showing no/weak homology to other protein sequences in a publicly available database (SWISSPROT) may be carried out by employing the following steps:

I. generating computationally overlapping peptide library from the protein sequences of unknown function,

II. sorting computationally the peptides of length 'N' (N is the length of the sliding window of amino acids) obtained as above, alphabetically, according to single letter amino acid code,

III. matching computationally the current library with peptide library of all functionally known proteins to obtain common peptides,

IV. locating computationally these common peptides in the original proteins and subsequently labeling them with their origin and location,

V. joining computationally the overlapping common peptides to obtain a long chain of invariant peptide sequences,

VI. assigning function to the unknown protein based on the function of the protein with which maximum length of peptide sequence identity is found. The more is the number of matches with the proteins of similar function the likelihood of functional assignment will be higher.

The particulars of the organisms such as their name, strain, accession number and other details are given below.

Genomes	Strain	Accession Number	Total Base Sequences	Date of Completion
Mycobacterium tuberculosis Cole,S.T., and et.al. *Nature* **393** (6685), 537-544 (1998)	H37Rv	AL123456	4411529 bp	Jun 11, 1998.
Bacillus subtilis Kunst,F. and et.al. *Nature* **390** (6657), 249-256 (1997)	DY	AL009126	4214814 bp	Nov 20, 1997
Mycoplasma genitalium Fraser,C.M., and et.al. *Science* **270** (5235), 397-403 (1995)	G37	L43967	580074 bp	Oct 30, 1995
Mycoplasma pneumonia Himmelreich,R., and et.al Nucleic Acids Res. **24** (22), 4420-4449 (1996)	M129	U00089	816394 bp	Nov 15, 1996
Escherichia coli Blattner,F.R.,. and et.al Science **277** (5331), 1453-1474 (1997)	K-12	U00096	4639221 bp	Oct 13, 1998.

Helicobacter pylori 26695 AE000511 1667867 bp Aug 6, 1997.
Tomb,J.-F., and et.al Nature **388** (6642), 539-547 (1997)

Haemophilus influenzae Rd L42023 1830138 bp Jul 25, 1995.
Fleischmann,R.D., and et.al Science **269** (5223), 496-512 (1995)

Genome	Proteins	Number of 8-mer peptides	No. of Proteins in which common peptides are found
Bacillus subtilis	4100	1174826	69
Escherichia coli	4289	1302149	81
Haemophilus influenzae	1709	504044	56
Helicobacter pylori	1566	474087	51
Mycoplasma genitalium	467	165523	30
Mycoplasma pneumonia	677	221216	43
Mycobacterium tuberculosis	3918	1252582	58

Brief description of the accompanying drawings:

Figure 1 shows a logic circuit of Peptide Library Matching Program.

Figure 2 shows a Logic circuit of Peptide Extraction Program.

Figure 3 shows a Logic circuit of Peptide Stitching Program.

Figure 4 shows crystal structures of three invariant peptides (**VRKRPGMYIG**, **LHAGGKFD** and **SGGLHGVG**) from DNA gyrase B protein.

The invention is explained with the help of the following examples and should not be construed to limit the scope of the present invention.

Examples

Example 1

1. The peptide library creation program (PEPLIB)

The purpose of the program is to create a non-redundant peptide library of user specified window length 'N' of a given genome by sliding the window by one amino acid residue at a time.

The program works as follows:

The internet downloaded FASTA format files obtained from http://www.ncbi.nlm.nih.gov were saved by the name <organism_name>.pep are passed as input to the PERL program which creates unique peptides of length as specified at the time of execution.

Input / Output file format:

Downloaded Files and their format:

<organism_name>. pep : file which stores the annotation & the protein sequence

<organism_name> refers to

Tb (*Mycobacterium tuberculosis*) **Bs** (*Bacillus subtilis*) **Mg** (*Mycoplasma genitalium*) **Mp** (*Mycoplasma pneumonia*) **Ec** (*Escherichia coli*) **Hp** (*Helicobacter pylori*) **Hi** (*Haemophilus influenzae*)

Format: FASTA

">gi|"<annotation>
<<the entire protein sequence.....

For example,

>gi|2808711|emb|CAA16238.1| dnaA
MTDDPGSGFTTVWNAVVSELNGDPKVDDGPSSDANLSAPLTPQQRAWLNLVQPLTIVEGF
ALLSVPSSFVQNEIERHLRAPITDALSRRLGHQIQLGVRIAPPATDEADDTTVPPSENPATTS
PDTTTDNDEIDDSAAARGDNQHSWP.....

>gi|3261513|emb|CAA16239.1| dnaN
MDAATTRVGLTDLTFRLLRESFADAVSWVAKNLPARPAVPVLSGVLLTGSDNGLTISGFD
YEVSAEAQVGAEIVSPGSVLVSGRLLSDITRALPNKPVDVHVEGNRVALTCGNARFSLPTM
PVEDYPTLPTLPEETGLLPAE.....

The output file: <organism_name><peptide_length>.txt

Format:

<all unique peptides of length specified at the time of execution>

for example format of Tb8.txt:

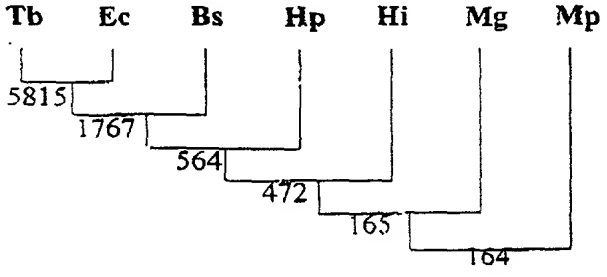
AAAAAAAA
AAAAAAAG
AAAAAAAQ
AAAAAAAS
AAAAAAAT
.....

Example 2

The peptide library matching program (PEPLIMP)

The purpose of the program is to compare the user defined peptide libraries with each other and report the common/ unique peptides. The output files of the program PEPLIB are used as input for the PEPLIMP program. As the program is executed the user is prompted to select the libraries that are to be compared. Depending upon the libraries selected an output file is generated having common peptides (Fig 1). Comparison of 8-mer peptide libraries of the above mentioned seven organisms resulted into 164 eight-mer peptides.

Comparison of four pathogenic organisms such as *Mycobacterium tuberculosis*, *Helicobacter pylori*, *Mycoplasma pneumonia* and *Haemophilus influenzae* resulted in 206 invariant peptides and comparison of three non-pathogenic organisms such as *Bacillus subtilis*, *Mycoplasma genitalium* and *Escherichia coli* resulted in 601 invariant peptides. The comparison tree looks like:



Example 3

The peptide extraction program (PEPXTRACT)

This program takes the output of PEPLIMP program i.e., all the invariant peptides as input and locates these peptides in the protein sequences from the original database and labels them with the protein identification number (PID), location and organism name for further analysis. The logic circuit of this program is explained in the flow chart shown in figure 2.

Example 4

The peptide stitching program (PEPSTITCH)

This program intelligently removes the overlapping invariant peptides and reports all the continuous stretch of invariant peptide present in the protein under consideration. This is done by first grouping the 'N'-mer peptides from the same protein of an organism and then keeping track on the their location they are merged into a long single peptide. The logic circuit of this program is shown in figure 3.

Example 5

Prediction of function of hypothetical protein

An invariant peptide having sequence **FSGGQRQR** was found to exist in oppF/dppF proteins of six organisms out of the seven examined (except for in *M. tuberculosis*). This protein functions as an ATP binding protein. Since this invariant peptide has also been found to be located on the hypothetical protein encoded by **Rv1273c** gene in *M. tuberculosis*, it is suggested that this protein encoded by **Rv1273c** gene must function as ATP binding protein as it holds the signature of this class of protein.

Example 6

Prediction of function of hypothetical protein

Another invariant peptide having sequence **GIVGLPNVGKS** was found in proteins having GTP binding function in six bacteria out of the seven examined (except for in *M. tuberculosis*) where as the same invariant sequence is present in hypothetical protein encoded by **Rv1112** gene in *M. tuberculosis*. It is strongly suggested that this hypothetical protein may have GTP binding property as it holds the signature of this class of protein.

Example 7

Drug target identification based on invariant peptide motifs

Enzyme DNA gyrase is known to reduce supercoiling of DNA. This protein is absent in human and has been considered as a potential drug target. However, the exact sequence to which the drug molecules should be targeted is not yet clear. The peptides such as **VRKRPGMYIG, LHAGGKFD, SGGLHGVG, LPGKLADC, VEGDSAGG** and **QRYKGLGEM** that are invariant across many pathogenic and non-pathogenic bacterial DNA gyrase beta subunit, but absent in host, are the structural determinants which could be used as potential drug targets against bacterial infections. The crystal structures of three of these peptides are shown in fig 4.

Example 8

Assignment of a function to a protein of unknown function

With the help of this method one can assign function to a protein of unknown function showing no/weak homology to other protein sequences in a publicly available database (SWISSPROT) by employing the following steps:

I. generating computationally overlapping peptide library from the protein sequences of unknown function,

II. sorting computationally the peptides of length 'N' (N is the length of the sliding window of amino acids) obtained as above, alphabetically, according to single letter amino acid code,

III. matching computationally the current library with peptide library of all functionally known proteins to obtain common peptides,

IV. locating computationally these common peptides in the original proteins and subsequently labeling them with their origin and location,

V. joining computationally the overlapping common peptides to obtain a long chain of invariant peptide sequences,

VI. assigning function to the unknown protein based on the function of the protein with which maximum length of peptide sequence identity is found. The more is the number of matches with the proteins of similar function the likelihood of functional assignment will be higher.

Advantages:

1. Main advantage of the present invention is to provide a new method of genome-wise comparison of large number (thousands) of proteins of one organism with proteins of other organisms simultaneously to arrive at invariant peptide sequence motif signatures.
2. It provides a rapid method of identification of invariant peptide motifs.
3. It provides a simple and highly accurate method of determining invariant peptide motifs as it does not involve any complex mathematical calculations.
4. It provides a basis for a screening assay for broad-spectrum antibacterial compounds.

References:

Altschul, S.F., Carol, R.J., & Lipman, D.J. (1990). Basic local alignment search tool. J.Mol.Biol. 215, 403-410.

Cutler N.S., Heitman J., Cardenas M.E., (1999). TOR kinase homologs function in a signal transduction pathway that is conserved from yeast to mammals. Mol Cell Endocrinol 155(1-2), 135-142.

Ghannoum, M.A. and Rice, L.B., (1999). Antifungal agents: mode of action, mechanisms of resistance, and correlation of these mechanisms with bacterial resistance. Clin Microbiol Rev 12(4), 501-517.

McCafferty D.G., Cudic, P., Yu, M.K., Behenna, D.C., Kruger, R., (1999). Synergy and duality in peptide antibiotic mechanisms. Curr Opin Chem Biol 3(6), 672-680.

Porse, B.T., & Garrette. R.A.(1999).Ribosomal mechanics, antibiotics, and GTP hydrolysis. Cell 97, 423-426.

Presenell, S.R., Cohen, B.I., & Cohen, F.E., (1992). A segment based approach to protein secondary structure prediction. Biochemistry 31, 983-993.

Rooman, M.J., & Wodak, S.J. (1988). Identification of predictive sequence motifs limited by protein structure database size. Nature 335, 45-49.

Wilbur, W.J., & Lipman, D.J. (1983). Rapid similarity searches of nucleic acid and protein data banks. Proc Natl Acad Sci USA 80, 726-730.

Wimberly, B.T., Guymon, R., McCutcheon, White, S.W., & Ramakrishnan, V., (1999). A etailed view of a ribosomal active site: The structure of the L11-RNA complex. Cell 97, 491-502.

Claims:

We claim.

1. A computer-based method for identifying invariant peptide motifs useful as drug targets wherein the said method comprises the steps of:

i) generating computationally overlapping peptide libraries from all the protein sequences of the selected organisms available at http://www.ncbi.nlm.nih.gov,

ii) sorting computationally the peptides of length 'N' obtained as above, alphabetically, according to single letter amino acid code,

iii) matching computationally common peptide sequences of the selected bacteria,

iv) locating computationally these common peptides in the original proteins and subsequently labeling them with their origin and location,

v) joining computationally the overlapping common peptides to obtain a long chain of invariant peptide sequences,

vi) annotating secondary structure of these conserved peptides from the crystal structure database,

vii) comparing pathogenic strain genomes against genomes of non-pathogenic strains and selecting the sequences not commonly conserved in these two groups,

viii) validating computationally the invariant sequence motifs as potential drug target sequence by searching for the given conserved sequences in the host genome and rejecting the ones present in the host genome.

2. The method of claim 1 wherein the length of the sliding window of length 'N' ranges from 4 to any length of amino acid residues.

3. The method of claim 1 wherein the protein sequence data is taken from any organism but not specifically limited to microbes such as *Mycoplasma pneumoniae*, *Helicobacter pylori*, *Hemophillus influenzae*, *Mycobacterium tuberculosis*, *Mycoplasma genitalium*, *Bacillus subtillis*, *Escherichia coli*.

4. A method as claimed in claim 1 where conserved peptide motifs as identified comprising:

1. AAQSIGEPGTQLT
2. AGDGTTTAT
3. AGRHGNKG
4. AHIDAGKTTT
5. CPIETPEG
6. DEPSIGLH
7. DEPTSALD
8. DEPTTALDVT
9. DHAGIATQ
10. DHPHGGGEG
11. DLGGGTFD
12. DVLDTWFSS
13. ERERGITI
14. ERGITITSAAT
15. ESRRIDNQLRGR
16. FSGGQRQR
17. GEPGVGKTA
18. GFDYLRDN
19. GHNLQEHS
20. GIDLGTTNS
21. GINLLREGLD
22. GIVGLPNVGKS
23. GKSSLLNA
24. GLTGRKIIVDTYG
25. GPPGTGKTLLA
26. GPPGVGKT
27. GSGKTTLL
28. GTRIFGPV
29. IDTPGHVDFT
30. IIAHIDHGKSTL
31. INGFGRIGR
32. IREGGRTVG
33. IVGESGSGKS
34. KFSTYATWWI
35. KMSKSKGN
36. KMSKSLGN
37. KNMITGAAQMDGAILVV
38. KPNSALRK
39. LFGGAGVGKTV
40. LGPSGCGK
41. LHAGGKFD
42. LIDEARTPLHSG
43. LLNRAPTLH
44. LPDKAIDLIDE
45. LPGKLADC
46. LSGGQQQR
47. MGHVDHGKT
48. NADFDGDQMAVH
49. NGAGKSTL
50. NLLGKRVD
51. NTDAEGRL
52. PSAVGYQPTLA
53. QRVAIARA
54. QRYKGLGEM
55. RDGLKPVHRR
56. SALDVSIQA
57. SGGLHGVG
58. SGSGKSSL
59. SGSGKSTL
60. SVFAGVGERTREGND
61. TGRTHQIRVH
62. TGVSGSGKS
63. TLSGGEAQRI
64. TNKYAEGYP
65. TPRSNPATY
66. VEGDSAGG
67. VRKRPGMYIG

5. A method as claimed in claim 1 wherein the number of invariant peptides varies according to the relatedness among the organisms and the number of organisms being compared.

6. A method as claimed in claim 1-4 wherein the invariant sequences belong to following proteins as available in the database http://www.ncbi.nlm.nih.gov wherein the said list of proteins comprise:

I DNA DIRECTED RNA POLYMERASE BETA CHAIN

II EXCINUCLEASE ABC SUBUNIT A

III EXCINUCLEASE ABC SUBUNIT B

IV DNA GYRASE SUBUNIT B

V ATP SYNTHASE BETA CHAIN

VI S-ADENOSYLMETHIONINE SYNTHETASE

VII GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE

VIII ELONGATION FACTOR G (EF-G)

IX ELONGATION FACTOR TU (EF-TU)

X 30S RIBOSOMAL PROTEIN S12

XI 50S RIBOSOMAL PROTEIN L12

XII 50S RIBOSOMAL PROTEIN L14

XIII VALYL tRNA SYNTHETASE (VALRS)

XIV CELL DIVISON PROTEIN FtSH HOMOLOG

XV DnaK PROTEIN (HSP70)

XVI GTP BINDING PROTEIN LepA

XVII TRANSPORTER

XVIII OLIGOPEPTIDE TRANSPORT ATP BINDING PROTEIN OPPF

7. A method as claimed in claim 1 wherein the said method of comparing the peptide libraries as given in step (iii) of claim 1 is carried out by following the steps given in figure 1.
8. A method as claimed in claim 1 wherein the said method of locating the common peptides in the original protein sequences as given in step (iv) of claim 1 is carried out by following the steps given in figure 2.
9. A method as claimed in claim 1 wherein the said method of creating a common peptide of variable length after removing the overlappings as given in step (v) of claim 1 is carried out by following the steps given in figure 3.
10. A microprocessor based system for performing the methods of the invention which comprises:

i) means of determining the amino acid sequence window for creation of peptide library and subsequent origin tagging,

ii) means of comparing the peptide library,

iii) locating computationally these common peptides in the original proteins and subsequently labeling them with their origin and location,

iv) joining computationally the overlapping common peptides to obtain a long chain of invariant peptide sequences,

11. A computer based system for performing the methods of the invention further comprising a central processing unit, executing peptide library creating program (PEPLIB), peptide library matching program (PEPLIMP), peptide stitching program (PEPSTITCH), peptide extraction program (PEPXTRACT) wherein the said programs are all stored in a memory device accessed by the central processing unit connected to a display on which the central processing unit displays the screens of the above mentioned programs in response to user inputs with a user interface device.

12. A method for assigning function to a protein of unknown function showing no/weak homology to other protein sequences in a publicly available database (SWISSPROT) by employing the following steps:

I. generating computationally overlapping peptide library from the protein sequences of unknown function,

II. sorting computationally the peptides of length 'N' (N is the length of the sliding window of amino acids) obtained as above, alphabetically, according to single letter amino acid code,

III. matching computationally the current library with peptide library of all functionally known proteins to obtain common peptides,

IV. locating computationally these common peptides in the original proteins and subsequently labeling them with their origin and location,

V. joining computationally the overlapping common peptides to obtain a long chain of invariant peptide sequences,

VI. assigning function to the unknown protein based on the function of the protein with which maximum length of peptide sequence identity is found. The more is the number of matches with the proteins of similar function the likelihood of functional assignment will be higher.

A computer based method for identifying peptides useful as drug targets

Abstract

The present invention relates to a novel computer based method for performing genome-wise comparison of several organisms, the said computational method involves creation of peptide libraries from protein sequences of several organisms and subsequent comparison leading to identification of conserved invariant peptide motifs, and to this end several invariant peptide motifs have been identified by direct sequence comparison between various bacterial organisms and host genomes without any *a priori* assumptions, and the present method is useful for identification of potential drug targets and can serve as drug screen for broad-spectrum antibacterials as well as for specific diagnosis of infections, and in addition, for assignment of function to proteins of yet unknown functions with the help of such invariant peptide motif signatures.

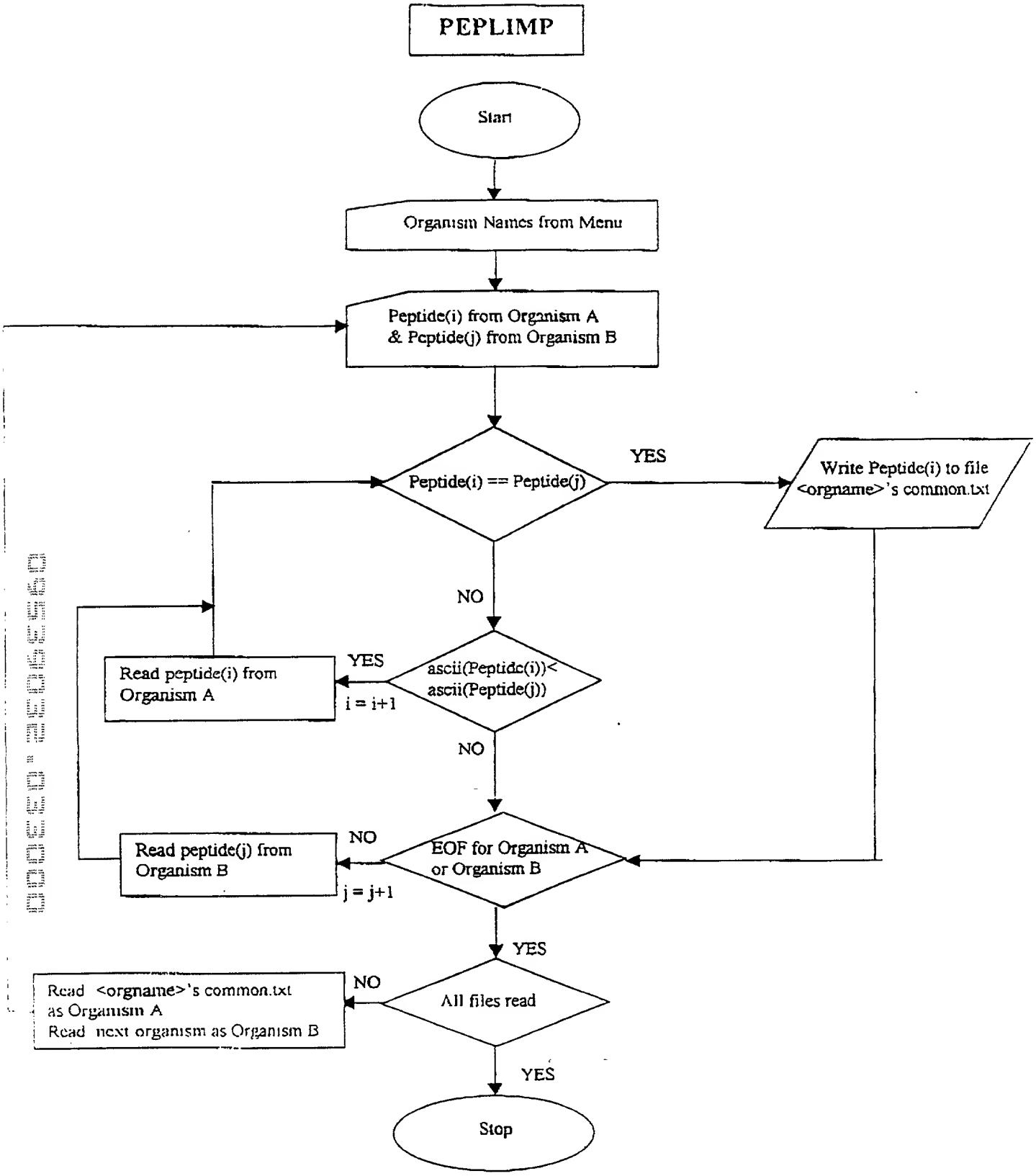

PEPLIMP
Start
Organism Names from Menu
Peptide(i) from Organism A
& Peptide(j) from Organism B
Peptide(i) == Peptide(j)
YES
Write Peptide(i) to file
<orgname>'s common.txt
NO
ascii(Peptide(i))<
ascii(Peptide(j))
YES
Read peptide(i) from
Organism A
i = i+1
NO
EOF for Organism A
or Organism B
NO
Read peptide(j) from
Organism B
j = j+1
YES
All files read
NO
Read <orgname>'s common.txt
as Organism A
Read next organism as Organism B
YES
Stop

FIG. 1

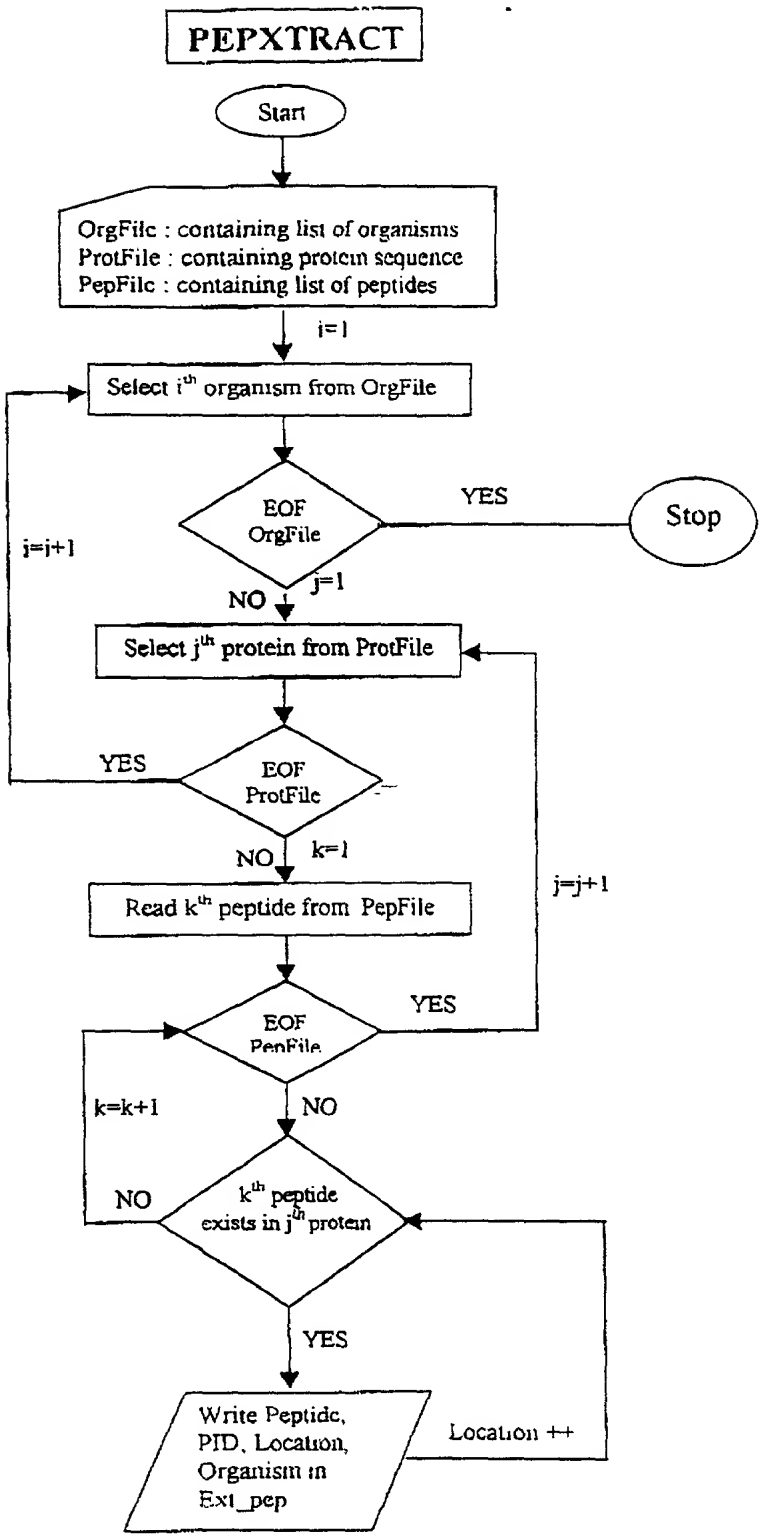

PEPXTRACT
Start
OrgFile : containing list of organisms
ProtFile : containing protein sequence
PepFile : containing list of peptides
i=1
Select i^th organism from OrgFile
EOF OrgFile
YES
Stop
j=j+1
j=1
NO
Select j^th protein from ProtFile
YES
EOF ProtFile
NO
k=1
Read k^th peptide from PepFile
j=j+1
EOF PepFile
YES
k=k+1
NO
NO
k^th peptide exists in j^th protein
YES
Write Peptide, PID, Location, Organism in Ext_pep
Location ++

FIG. 2

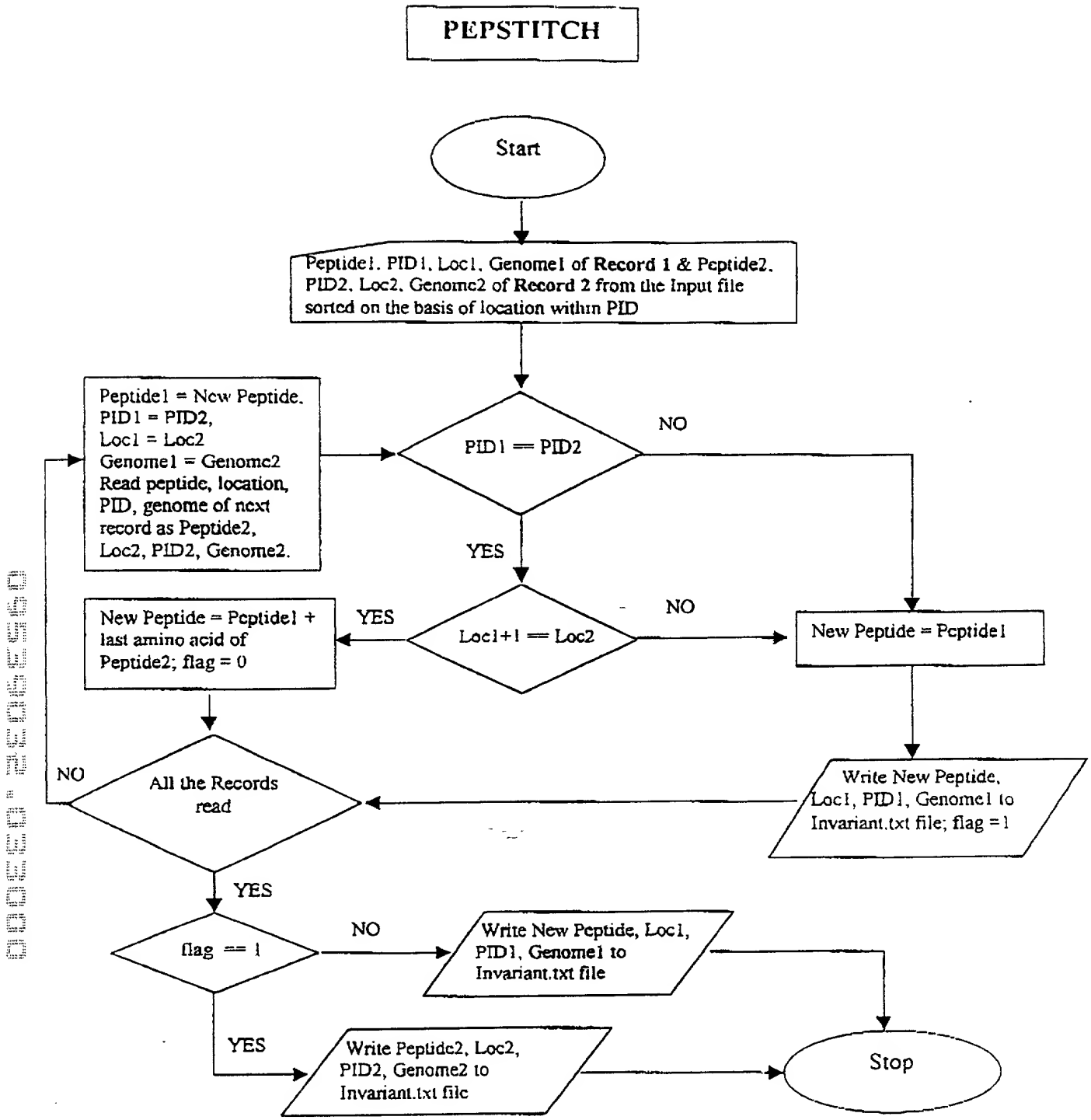

PEPSTITCH
Start
Peptide1, PID1, Loc1, Genome1 of Record 1 & Peptide2, PID2, Loc2, Genome2 of Record 2 from the Input file sorted on the basis of location within PID
Peptide1 = New Peptide, PID1 = PID2, Loc1 = Loc2 Genome1 = Genome2 Read peptide, location, PID, genome of next record as Peptide2, Loc2, PID2, Genome2.
PID1 == PID2
NO
YES
Loc1+1 == Loc2
YES
NO
New Peptide = Peptide1 + last amino acid of Peptide2; flag = 0
New Peptide = Peptide1
Write New Peptide, Loc1, PID1, Genome1 to Invariant.txt file; flag = 1
All the Records read
NO
YES
flag == 1
NO
Write New Peptide, Loc1, PID1, Genome1 to Invariant.txt file
YES
Write Peptide2, Loc2, PID2, Genome2 to Invariant.txt file
Stop

FIG. 3

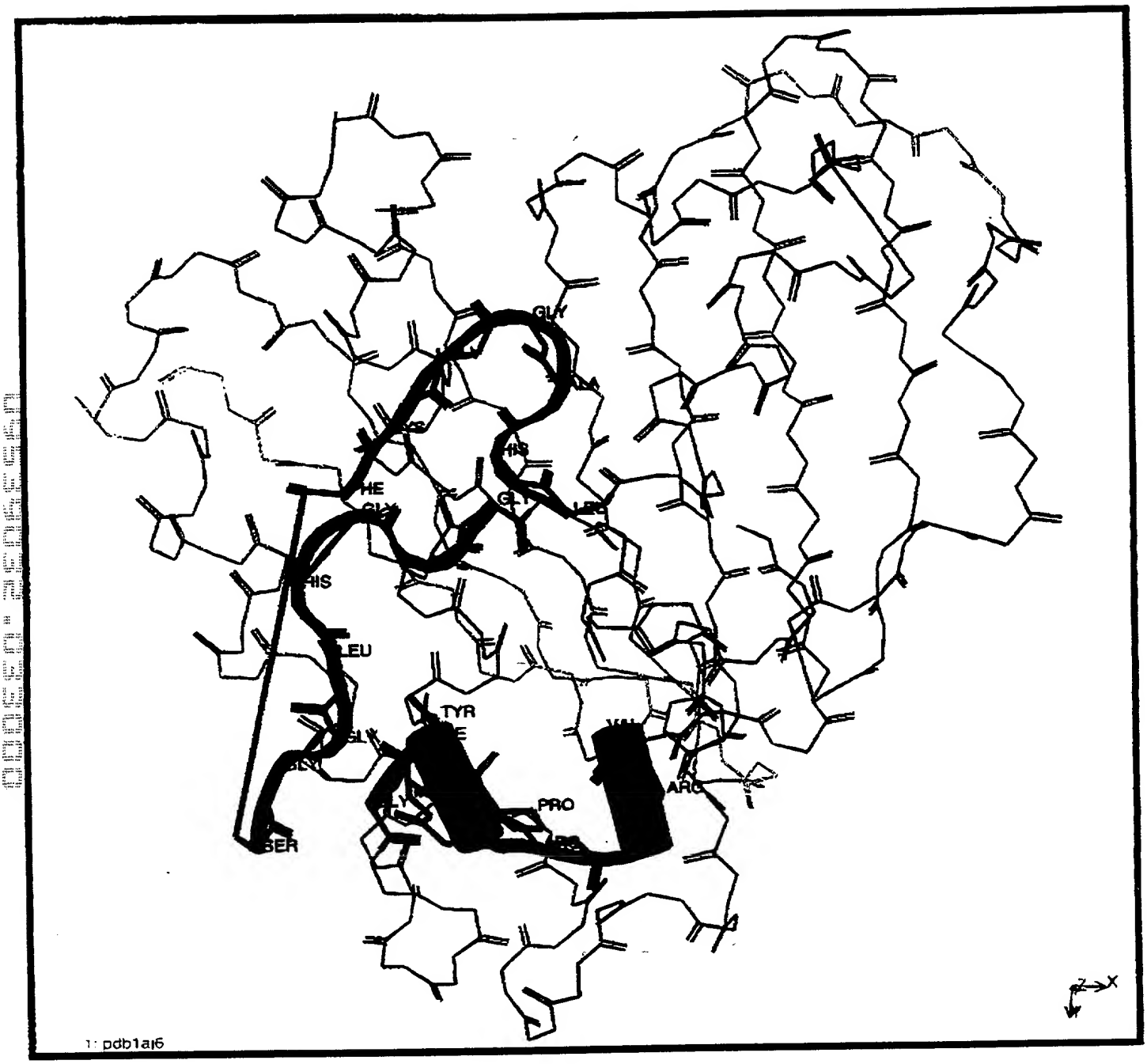


FIG. 4

Attorney's Docket No.: 07064-010001
Client's Ref. No.: 0317 - NF

COMBINED DECLARATION AND POWER OF ATTORNEY

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name,

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled A COMPUTER BASED METHOD FOR IDENTIFYING PEPTIDES USEFUL AS DRUG TARGETS, the specification of which:

[X] is attached hereto.
[] was filed on _ as Application Serial No. _ and was amended on _____.
[] was described and claimed in PCT International Application No. _____ filed on _____ and as amended under PCT Article 19 on _____.

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose all information I know to be material to patentability in accordance with Title 37, Code of Federal Regulations, §1.56.

I hereby appoint the following attorneys and/or agents to prosecute this application and to transact all business in the Patent and Trademark Office connected therewith:

John W. Freeman, Reg. No. 29,066
Anita L. Meiklejohn, Reg. No. 35,283
Timothy A. French, Reg. No. 30,175
Eldora L. Ellison, Reg. No. 39,967

Address all telephone calls to JOHN W. FREEMAN, ESQ. at telephone number (617) 542-5070.

Address all correspondence to JOHN W. FREEMAN, ESQ. at:

FISH & RICHARDSON P.C.
225 Franklin Street
Boston, MA 02110-2804

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patents issued thereon.

Full Name of Inventor: Samir Kumar Brahmachari

Inventor's Signature: _____ Date: _____
Residence Address: Central for Biochemical Technology, Mall Road, Delhi INDIA 110 007
Citizenship: India
Post Office Address: Central for Biochemical Technology, Mall Road, Delhi INDIA 110 007

Combined Declaration and Power of Attorney

Full Name of Inventor: Debasis Dash

Inventor's Signature: _____ Date: _____

Residence Address: Central for Biochemical Technology, Mall Road, Delhi INDIA 110 007

Citizenship: India

Post Office Address: Central for Biochemical Technology, Mall Road, Delhi INDIA 110 007